DOCUMENT RESUME

ED 365 689 TM 020 835

AUTHOR	Kane, Michael
TITLE	The Validity of Performance Standards.
SPONS AGENCY	National Assessment Governing Board, Washington, DC.
PUB DATE	19 Jun 93
NOTE	46p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Standards; *Achievement Tests; *Cutting Scores; Data Collection; Educational Policy; Elementary Secondary Education; Higher Education; *Performance; Research Methodology; Test Use; *Validity
IDENTIFIERS	*High Stakes Tests; *Performance Based Evaluation

ABSTRACT

A general framework is provided for examining the validity of performance standards for high-stakes achievement tests. The emphasis is on conceptual issues and broadly defined methodological questions, the types of validity evidence that can be collected, and the advantages and limitations of different types of evidence. Validation consists of a demonstration that the proposed passing score can be interpreted as representing the level of achievement specified in the proposed performance standard. The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version of the desired level of competence. The analysis addresses the question of the arbitrariness of the passing score by identifying two assumptions that are involved in adopting a passing score. The first is that it corresponds to a specified performance standard, and the second is that the specified standard is appropriate. Support for the first, descriptive, assumption is to be derived mainly from procedural evidence and internal validity checks. Support for the policy assumption is to be derived mainly from procedural evidence and external validity checks. Even if all available checks on the validity of the standard are implemented, the best that can be done is to show that the proposed standard is reasonable or acceptable. (Contains 79 references.) (SLD)

The Validity of Performance Standards

Michael Kane
Department of Kinesiology
University of Wisconsin-Madison

This paper was prepared for the National Assessment Governing Board. The findings and opinions expressed in this paper are those of the author and do not necessarily represent the position of NAGB.

6/19/93

The intent of this paper is to provide a general framework for examining the validity of performance standards for high-stakes achievement tests. The empnasis is on conceptual issues and broadly defined methodological questions, on the kinds of validity evidence that can be collected, and on the advantages and limitations of different types of evidence.

In developing this framework, two fundamental questions had to be addressed. First, there is the question of what we mean by the validity of a performance standard? In psychometrics, usage of the term "validity" assumes that we have a score scale based on some assessment procedure, and we have a proposed interpretation for examinees' scores. The question of validity asks whether the proposed interpretation is legitimate. So, validity is a property of the interpretation assigned to test scores and not a property of the test itself or of the test scores (Messick, 1989). What are we validating when we validate a performance standard? That is, what is it that we are interpreting (i.e., the analog of the score scale) and what is the interpretation (i.e., the analog of the proposed interpretation of the score scale)?

In responding to this question, it is useful to draw a distinction between the passing score, defined as a point on the score scale, and the performance standard, defined as the minimally adequate level of performance for some purpose. Validation then consists of a demonstration that the proposed passing score can be interpreted as representing the level of achievement specified in the proposed performance standard. The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version of the desired level of competence. In this paper, the term "standard" will be used to refer to the desired level of competence, without distinguishing between the passing score and the performance standard. In much of the literature on standard setting, the distinction between the passing score and the performance standard is not explicitly drawn, making it difficult to evaluate the validity of a proposed interpretation for the passing score. Maintaining a clear distinction between the passing score and the corresponding performance standard helps to avoid such confusion.

The second question involves what Glass (1978) has called the "arbitrariness" of passing scores. Clearly, there is an element of arbitrariness in all standard setting, but most standards are not completely arbitrary (Scriven, 1978; Popham, 1978). Some standards seem quite arbitrary; the tradition of requiring 70% correct on some tests seems especially arbitrary, because we know that for any group of examinees, we can probably make the items easy enough so that everyone gets more than 70% correct or difficult enough so that nobody gets more than 70% correct. Some standards do not seem at all arbitrary; a requirement that a lifeguard be able to swim a certain distance in a certain time and then swim back pulling a struggling victim does not seem particularly arbitrary. If we are to make much sense out of the arbitrariness question, we need some understanding of the source of arbitrariness in performance standards and passing scores, of why some standards seem so much more arbitrary than others, and of how we might control or limit the arbitrariness in our standards.

The analysis presented here addresses the question of arbitrariness by identifying two assumptions that are involved in adopting a passing score and its

associated performance standard. The first assumption claims that the passing score corresponds to a specified performance standard, which is defined in terms of level of achievement or skill in some area. The evaluation of this first claim does not get us into any more arbitrariness than we usually encounter in validation efforts. In fact, if we have validated the score scale well enough so that we know what the different scores mean in terms of levels of achievement, presumably we know what the passing score means as a level of achievement. So, the first assumption does not present any special problems of "arbitrariness", other than the penumbra of vagueness in all of our constructs and the uncertainty introduced by errors of measurement.

The second assumption claims that the specified standard is appropriate, that is, the level of performance is just high enough to accomplish the purpose for which the decision process was implemented. This second assumption involves the adoption of a policy, explicitly or implicitly, and the values and expectations about consequences inherent in the policy are likely to be subject to dispute. As Jaeger (1990, p. 18) puts it:

> No conventional test validation procedure will provide evidence that any score-based dichotomization of the ability scale into two categories labeled "competent" and "incompetent" is correct. We know that the dichotomization is judgmentally based, arbitrary, and, wherever placed on the ability scale, will not result in reliable differences in distributions of performance on any valued criterion for groups adjacent to the point of dichotomization.

The essential, unavoidable arbitrariness in standard setting is found in the details of the second assumption. It is the arbitrariness that always exists in social and political policy decisions. These decisions could be changed, and often are changed, when made by different persons, at different times, or under different circumstances. That is, the passing score could always be moved up or down a bit without violating any fundamental principles; the final choice is a matter of judgment.

The standard of performance for lifeguards does not seem arbitrary because the policy decisions inherent in this standard (e.g., that one of the maim functions of a lifeguard is to rescue people who are drowning) are not under dispute, and this general policy and the context (i.e., the typical circumstance in which rescues occur) determine the specific performance standards within fairly narrow limits. The passing scores that seem more arbitrary are those that are not based on an accepted policy. What we should expect of every high school graduate in mathematics, science, art, etc. is not so clear.

If the policy decisions have already been made, standard setting can proceed in a systematic way, because the only questions to be answered by the standard-setting study are the technical issues of implementation which are associated with the first assumption. If the policy decisions have yet to be made, all of the arbitrariness inherent in the making of social/political decisions, including the tradeoffs among competing values and goals, enter the picture.

This paper examines various approaches to the validation of performance standards and, in doing so, addresses some issues involved in implementing standard-setting procedures. Before embarking on any standard-setting method, however, it is important to consider the fundamental issue of whether it is necessary or useful to employ a passing score, or multiple cutoff scores, in a particular case (Burton, 1978; Glass, 1978; Levin, 1978). The use of specific performance standards in interpreting assessment results always involves some risks (Levin, 1978), which should be weighed against potential gains. Assuming that it is necessary or useful to employ a passing score, it is important to be clear about what we want to achieve in making pass/fail decisions, so that our goals can guide our choices at various stages in the standards-setting process.

Standard-Setting Methods

In addition to the descriptions found in the original sources (e.g., Angoff, 1971; Ebel, 1972; Jaeger, 1982; Nedelsky, 1954), the commonly used standard-setting methods have been clearly described in a number of reviews (e.g., Livingston and Zieky, 1982, 1983; Berk, 1986; Shepard, 1979, 1980, 1984; Jaeger, 1989), and therefore do not need to be described in detail here. Brief descriptions of the most commonly used methods are provided for ease of reference.

Meskauskas (1976) drew the distinction between state models, which assume that each examinee is in one of two possible states, competent or not competent, and continuum models, which assume that the attribute being measured is a continuous variable. State models have not been used much in practice, because most practical standard-setting problems require that a passing score be set on a score scale with a range of possible values. Therefore, in this discussion of the validity of performance standards, I will focus on continuum models.

Berk (1986) lists a number of methods that have been proposed for adjusting passing scores in various ways once an initial "true" passing score is established. Most of these methods seek to define a passing score on the observed score scale that minimizes false positive decisions, false negative decisions, or both (Hambleton and Novick, 1973; Hambleton, Swaminathan, Algina, and Coulson, 1978; van der Linden and Mellenberg, 1977). This paper focuses on the prior issue of setting and validating standards in situations where there are no existing standards, and shall not discuss methods for subsequently adjusting the passing score.

Test-centered Models

The most commonly used methods for standard setting are those that Jaeger (1989) refers to as the "test-centered models". In the test-centered models, judges set the standard by reviewing the items included in the test and deciding on the level of performance on these items that will be considered just adequate. There are a number of ways for the judges to do this.

In the basic form of the Angoff (1971) procedure, the judges are asked to envision a minimally competent examinee and to estimate the probability that this

examinee will answer each item correctly. This probability is called a minimum pass level, or MPL. To make things easier, the judges are sometimes told to imagine a group of 100 minimally competent individuals and to estimate the number or proportion of these individuals who would answer the item correctly. In either case, the MPLs are averaged over judges to get the item MPL, and the item MPLs are summed over the items in the test to get a passing score. There are many variations on the Angoff procedure; Berk (1986) lists eight variations.

Ebel's (1972) procedure has the judges categorize the items in a test along two dimensions, according to their difficulty and their relevance to the decision to be made. With three levels of difficulty and four levels of relevance, there would be 12 categories. The judges then decide on the proportion of items in each category that a borderline examinee would answer correctly. The proposed passing score is computed by multiplying the number of items in each category by the proportion of items in the category that would be answered correctly by a borderline candidate, and then summing over the twelve categories.

The Nedelsky (1954) procedure was specifically designed for multiple-choice items. For each item, the judges decide on how many of the response options a minimally competent examinee would recognize as being incorrect. The minimal pass level for the item is computed as one over the number of options remaining after the obviously incorrect options are removed from consideration. The Nedelsky procedure can be thought of as assuming that the minimally competent examinee responds by eliminating those options recognized as being wrong and guessing among the remaining options. So, on a five-option item, if a judge decides that the minimally competent examinee would recognize two options as being wrong, the Nedelsky MPL would be 1/3 or 0.33, because three options remain after the two obviously wrong options are eliminated. The MPLs are summed over items to get the passing score on the test.

Jaeger's method (1982, 1989) differs from most other test-centered methods in a number of important ways, all of which tend to emphasize the role of standard setting as the development of policy rather than as a technical problem of parameter estimation.

First, the Jaeger (1989, p. 494) procedure defines the ideal situation as "sampling all populations that have a legitimate interest in the outcomes of competency testing". The political nature of many standard setting issues is explicitly recognized, and the need to consider alternative points of view is emphasized. Most of the other methods have traditionally relied on a single panel of expert judges.

Second, Jaeger (1989, p. 494) would ask judges to decide, yes or no, whether every examinee who passes the examination should be able to answer each question. The focus is clearly on <u>setting</u> the standard of what passing examinees <u>should</u> be able to do, rather than on estimating a parameter for a hypothetical population of minimally competent examinees.

Third, the Jaeger procedure is iterative. The process goes through several iterations, and, at each stage, the judges are given feedback on examinee performance and/or on the collective judgments of various groups of judges. Therefore, the Jaeger procedure builds some of the external and internal checks

on validity discussed later in this paper, into the standard-setting process. Note that, although the Angoff procedure did not originally involve an iterative approach, current applications of the Angoff procedure can, and usually do, involve two or more iterations, with judges getting feedback and reconsidering their judgments at each stage of the process.

Examinee-centered Models

In the examinee-centered models, the judges make pass/fail decisions about examinees. The passing score is then set by identifying a point on the score scale that would be consistent with these decisions.

In the borderline-group method (Livingston and Zieky, 1982) judges are required to identify individual test-takers as borderline, in the sense that their level of achievement is right around the performance standard. The judges, who may be teachers, supervisors, etc., use their experience with the individuals or some assessment other than the test to identify a group of borderline individuals. The median score for this group of borderline examinees is then used as the passing score. Livingston and Zieky (1982, p. 34) suggest an internal check on the procedure: if the scores of the borderline group are clustered together, the method is probably working well; if the scores are spread over a wide range, then the method is not working well.

In the contrasting-groups method (Livingston and Zieky, 1982), the judges categorize a group of examinees into two groups, those judged to be competent and those judged to be not competent. As in the borderline-group method, these judgements are made on some basis other than the test scores. A common suggestion is to have teachers, supervisors, or someone else who has experience with the examinee's performance make these judgments. After the score distributions for these two groups have been determined, the passing score is chosen so that it discriminates as well as possible between the competent group and the not-competent group.

The six methods outlined above are the most widely discussed standard-setting methods, and also seem to be the most widely used methods. My experience and an informal survey or organizations involved in standard setting indicates that the Angoff method is by far the most popular method for determining the passing scores to be used in high-stakes educational assessments and in licensure and certification examinations. Therefore, in discussing the kinds of analyses that might be used to evaluate the validity of performance standards, I will, for convenience, tend to speak as if the proposed standards were set with the Angoff procedure, and that the other methods (e.g., the borderline-group method) are available as a potential empirical check on the Angoff method. In all such cases, these roles could be reversed, with some other method providing the proposed passing score and the Angoff passing score employed as a check on validity.

Finally, it is worth noting a general approach to standard setting that has gotten less attention than it probably deserves. To the extent that the pass/fail distinction is being emphasized, it would seem to make sense to design the assessment procedure so that it yields high precision around the passing

score. That is, rather than apply the standard-setting procedures to an existing test, we would specify the performance standard and then develop the test to fit the standard.

Validity

Validity is a property of the interpretation assigned to test scores. The test itself is not validated, and test scores per se are not validated. The question of validity does not arise until we consider interpretations, and a proposed interpretation is valid if it is supported by appropriate evidence (AERA, APA, NCME, 1985; Cronbach, 1971; Jaeger, 1979; Linn, 1979; Madaus, 1983; Messick, 1989).

The interpretation of the score scale includes all of the statements to be made about examinees (or other objects of measurement, such as schools, classes, etc.) on the basis of the scores resulting from the assessment procedure, and all decisions about examinees based on the scores (Messick, 1981, 1988, 1989; Guion, 1974). It also includes the intermediate steps involved in getting from the score to the final conclusions and decisions about an examinee.

The dependence of validity on the details of the proposed interpretation is critical. In order to evaluate the plausibility of an interpretation, we have to be clear about what it claims. Every assessment procedure has some interpretations that are likely to be considered plausible, or valid, and other interpretations that are likely to be considered highly implausible. For example, suppose a test, consisting of 200 factual questions about various aspects of American History, is administered to tenth graders. A good case might be made for the interpretation of the resulting scores as measures of the students' knowledge of the factual content of American History. This interpretation is fairly plausible, especially if the questions have been selected in a systematic and sensible way, the items are clearly written, etc. The interpretation of the scores as indicators of the examinees' ability to analyze historical events in terms of their possible causes and consequences is far less plausible. The second of the two interpretations would require much more evidence to support its validity than the first. For any test, some interpretations are likely to be more plausible, for a given level of evidential support, than other interpretations.

The validity of the interpretation also depends on the population in which the assessment will be used and the context in which the assessment procedure is applied. An interpretation of test scores that is valid for one population, e.g., native speakers of English, may not be valid in another population, e.g., examinees for whom English is a second language or examinees with impaired vision or hearing. The circumstances in which the assessment is conducted may also have a major impact on the interpretation of scores. If the test were administered with severe time limits, it might function more as a test of reading speed than as a test of knowledge of history.

Interpretive Arguments

In order to evaluate the plausibility of an interpretation for test scores, it is necessary to be clear about what the interpretation claims, and one way to achieve greater precision in stating the interpretation is to lay it out in the form of an interpretive argument (Kane, 1992). The interpretive argument would specify the network of inferences leading from the score to the conclusions drawn about examinees and the decisions made about examinees, as well as the assumptions that support these inferences. The interpretive argument is intended to describe the reasoning involved in interpreting the scores in a particular way.

The interpretive argument provides a framework for developing validity evidence. One specifies the inferences and assumptions leading from the test scores to the statements and decisions included in the interpretation, identifies potential competing interpretations, and seeks evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counterarguments. To validate the interpretation is to support the plausibility of the corresponding interpretive argument with appropriate evidence.

An effective evaluation of the interpretive argument requires an investigation of key assumptions in the argument. It is not possible to prove all of the assumptions in interpretive arguments, but it is usually possible to develop some empirical evidence for or against doubtful assumptions. Because no particular piece of evidence is likely to be decisive, several types of evidence may be used to evaluate an assumption. The plausibility of an assumption is evaluated in terms of all of the available evidence.

Interpretive arguments are practical arguments, rather than formal (i.e., logical or mathematical) arguments and therefore cannot be proven. Even the simplest interpretive arguments contains many assumptions that cannot be taken for granted. In the example given earlier, involving the interpretation of history test scores as measures of knowledge of history, we assume that the examinees are motivated, that they have adequate time to complete the examination, that they have not been coached on the specific content of the items, etc. In most cases, we cannot prove all of the assumptions, and therefore, we cannot prove the interpretive argument. The best that we can do is to show that the interpretive argument is highly plausible by stating it clearly so that we know what it claims and what it assumes, by making sure that it is coherent in the sense that the conclusions follow from the assumptions, and by showing that the assumptions are reasonable. Parallel lines of evidence should be developed whenever this is possible, and plausible counter arguments should be considered.

The details of the interpretive argument will depend on the specific interpretation being proposed, the population to which the interpretation is applied, the specific data collection procedures being used, and the context in which measurement occurs. The particular mix of evidence needed to support the interpretative argument will be different for each case, but in each case, the aim of validation is to support the plausibility of the interpretative argument with appropriate evidence (Kane, 1992).

Validating Performance Standards

There is an important class of interpretations that involve decisions about whether examinees have met some standard of performance in some area of achievement. To validate the performance standard is to validate the interpretive argument in which this standard is used.

The decisions are made with some goal or purpose in mind, i.e., to ensure that passing examinees are ready for some activity or responsibility. Given the goal of the decision process, a relevant area of achievement is identified, an assessment procedure is developed (or chosen) to assess this area of achievement, and a passing score is set on the score scale for the measurement procedure. The presumption is that individuals with scores above the passing score have met some performance standard related to the goal of the decision process and individuals with scores below the passing score have not met the performance standard. So, we have a score scale that is interpreted in terms of level of achievement in some area, and we have a passing score that is interpreted in terms of a specific level of achievement, the performance standard. By requiring that examinees achieve a score at or above the passing score, we expect to ensure that they have met the corresponding performance standard, and thereby, we expect to achieve our goal.

Just as we do not validate a test but rather the interpretation assigned to test scores, we do not validate a passing score or a performance standard per se. Employing the approach outlined in the last section, we would evaluate the plausibility of the inferences made about examinees using the passing score and its associated performance standard.

Passing Scores and Performance Standards

In discussing the issues of validity associated with standard setting, it is important to be clear about the distinction between the "passing score" and the "performance standard". The particular point on the score scale that is used operationally to make decisions is the passing score. Examinees with scores at or above the passing score pass, and examinees with scores below the passing score fail. The performance standard represents the minimal acceptable level of achievement in some area, or on some type of activity, or some domain of related activities. The passing score is a point on the score scale, and the performance standard is a conceptual boundary between acceptable levels of achievement and unacceptable levels of achievement, or between being competent in an area and not being competent in the area. The passing score is a number, and the performance standard is a construct.

Assuming that the passing score is not selected in a completely arbitrary manner, it must be chosen to represent some intended distinction between passing and failing examinees. If the passing score is being used to select individuals for an educational program, job, etc., the intended distinction is likely to be between those who are well enough prepared to succeed in the educational program, job, etc. and those who are not prepared well enough for the program. If the pass/fail decision is made at the end of

a course, the intended distinction might be between those who have learned most of the content of the course and those who have learned relatively little.

In most cases, the pass/fail decision is intended to identify passing examinees who have several related characteristics, including perhaps mastery of some content domain, plus skill in performing certain tasks, which together make them ready to function effectively in some context. Failing examinees are judged to be unprepared to function effectively because they lack some or all of these requirements. The cluster of intended distinctions between passing and failing examinees included in the performance standard provides the basic interpretation assigned to the passing score. Given this model, the aim of a standard-setting study is to identify a passing score that achieves the goal of the decision process. As part of this standard-setting process, it is necessary to define a performance standard, which specifies the level of achievement needed to achieve the goal. The aim of the validation effort is to provide convincing evidence that the passing score does represent the intended performance standard and that this performance standard is appropriate, given the goals of the decision process.

The use of a passing score adds an explicit decision rule involving a passing score to the basic interpretation of the score scale. The decision rule categorizes all examinees with scores at or above the passing score as passing and examinees with scores below the passing score as failing. This decision process adds an additional layer of interpretation to the original interpretation of the score scale, to the effect that examinees with scores above the passing score are competent and those with scores below the passing score are not competent.

It is worth noting that the introduction of the performance standard also drops certain kinds of information from the interpretation, or at least deemphasizes this information. Typically, the interpretation of the basic score scale gives roughly equal emphasis to differences in scores, wherever they happen to fall on the scale. To the extent that we emphasize the results of pass/fail decisions, we give substantial attention to differences between passing and failing scores and relatively little attention to other differences.

Arbitrariness

Several authors have commented on the arbitrariness of performance standards. Glass (1978) pointed out the element of arbitrariness in performance standards and suggested that they not be used in most cases. In response, Popham (1978), Block (1978), Hambleton(1978), and Linn (1978) have suggested that although passing scores are arbitrary in the sense that they are based on judgment, they do not have to be arbitrary in the sense of being capricious.

One source of arbitrariness arises from the fact that the scale of achievement is conceived of as being continuous, and generally the benefits associated with achievement are assumed to be a monotone increasing function

of level of achievement. As achievement increases, the benefits associated with achievement increase. As a result, there is no simple and obvious way to choose a particular point on the score scale as the passing score. So, the choice of the passing score is arbitrary in the sense that there is no compelling reason why it could not be set a little higher or a little lower. As noted by Jaeger (1990), Shepard (1980) and others, examinees just below the passing score do not differ substantially from examinees just above the passing score.

The fact that the proposed performance standard and the associated passing score results from a policy decision rather than an entirely objective process of parameter estimation, can also be viewed as a source of arbitrariness. Policy decisions involve the integration of values with predictions (or guesses) about the consequences of various choices. As a result, individuals and groups may have serious differences of opinion that cannot be easily resolved. In particular, the different stakeholders in a decision process may have reasonable differences of opinions, based on different assumptions and values, about how high standards should be. As Werner (1978, p. 2) has suggested in a discussion of licensure examinations, any passing score can be criticized by persons whose views on what constitutes minimal acceptable competence for an occupation differ from the views that most significantly influenced the choice of a passing score. Any policy decision is arbitrary in the sense that it reflects a certain set of values and beliefs and not some other set of values or beliefs.

An Example

As we proceed in this section and subsequent sections, it will be useful to have an example to refer to.

Let's assume that we have a test of achievement in mathematics, covering topics in arithmetic, pre-algebra, algebra, and analytic geometry and calculus. This is a fairly broad range of topics for a single test, but it is not totally unrealistic, and it will facilitate the discussion of different levels of achievement.

I will assume that the test consists of a number of separately scoreable tasks or items, which may be either extended response or objective. I will assume that different items deal with different areas of content, i.e., that there are "arithmetic items", "algebra items", etc., and that the arithmetic items deal primarily with arithmetic, the algebra items deal primarily with algebra, etc.

I will assume further that the test has been constructed so that tasks dealing with the more advanced topics are generally more difficult than tasks dealing with the more elementary topics. As a result, an examinee with a low score is likely to answer some arithmetic items correctly, but is not likely to answer many of the algebra or calculus items correctly. (Note that this assumption is not true for many tests, because some traditional norm-referenced test development procedures are designed to create items that are fairly homogeneous in their empirical difficulty levels).

I have set up this example so that it is relatively straightforward. In actual practice, the situation is often far murkier. As noted above, it may not be easy to associate types of items or areas of content with regions on the score scale. We may have concerns about the validity of the score scale, about the potential for bias, or about the appropriateness of the test for a given purpose. Nevertheless, in discussing the implications of various kinds of evidence for the validity of standards, it is useful to keep the example as simple as possible.

The Assumptions Supporting the Decision

The interpretive argument can be thought of as going from an examinee's score to conclusions about the examinee's standing on the scale, and then to a pass/fail decision based on the passing score. Passing the test is interpreted as an indication that the examinee has reached the level of achievement specified in the performance standard corresponding to the passing score. The validity of the standard can be evaluated in terms of the plausibility of the assumptions supporting the inference that an examinee's level of achievement is adequate if and only if the examinee's score is at or above the passing score.

This inference is reasonable if the passing score is appropriate given the purpose of the decision process and is not reasonable if the passing score is not appropriate. In practice, the reasoning supporting the appropriateness of the passing score usually involves at least two assumptions. The first assumption, which I will refer to as the descriptive assumption, claims that the passing score corresponds to a specified performance standard, in the sense that examinees with scores above the passing score are likely to meet the standard, and examinees with scores below the passing score are not likely to meet the standard. The second assumption, which I will refer to as the policy assumption, claims that this performance standard is appropriate, given the purpose of the decision.

If we were using our math test to select students for a special science course, we might decide to set the passing score at a point on the score scale such that examinees with scores above the passing score can generally do the algebra items in the test and are therefore considered to be reasonably competent in algebra, and examinees with scores below the passing score generally cannot solve the algebra items and are therefore not considered competent in algebra. This performance standard is clearly defined. It could be made more precise by specifying what we mean by the expression "can generally do the algebra items on the test" (i.e., 80% correct), but the general intent of the standard is fairly clear.

The appropriateness of the performance standard and its associated passing score would depend to a large extent on how the science course is designed. If instruction in the course assumes a working knowledge of algebra and nothing more, the passing score would be appropriate. If the course does not involve mathematics to any appreciable extent, the standard may be far too high. If the course requires calculus, the standard is too low. So, it is largely a matter of policy (i.e., how we choose to design the course) whether

the requirement for a working knowledge of algebra is an appropriate standard. The descriptive assumption is satisfied if the pass/fail decision does reflect the difference between competence and lack of competence in algebra. The policy assumption is satisfied if this standard is appropriate given the purpose of the decision process.

There are many cases in which the interpretation of the decisions being made is in terms of "readiness" for something: for a course, for college, for professional practice, for a job, etc. To the extent that the requirements entailed by "readiness" are clearly defined, the appropriate performance standard is clear, and the standard setting effort can focus on the descriptive assumption by establishing a connection between the passing score and the performance standard associated with "readiness".

In other cases, the policy questions may be particularly salient. In some cases, the levels of achievement associated with different points on the score scale are clearly defined (e.g., for a fitness test in which the score is the number of pull-ups that can be done in a minute), and the problem is to decide on the performance level that would be most appropriate to use for a given purpose.

However, in almost all cases, we need to pay some attention to both assumptions. We choose a particular passing score because it corresponds to a performance standard, and we focus on a particular performance standard because we think that examinees who meet that standard can generally perform adequately in a particular context. Nevertheless, the distinction between these two assumptions is useful in analyzing how different kinds of evidence relate to the appropriateness of passing scores.

Each of our two assumptions involve arbitrariness. In most of the applications in which standard-setting procedures are used, the relationship between various performances, competencies, etc. that might be included in the performance standard and the score scale is assumed to be a relatively smooth, increasing function of test scores (i.e., performance improves gradually as a function of score, with no dramatic differences from one score point to the next). Therefore, there is no natural break point, at which to place the passing score. Performances do not change much if we move up or down a point or two on the score scale. There is, therefore, no way of deciding precisely what the passing score should be, because the performance standard is not precisely defined. And the more general and/or vague the performance standard, the more ambiguity there is in the passing score. Nevertheless, if a relatively clear performance standard is adopted (e.g., competence in algebra), it may be possible to make a good case that a passing score represents this performance standard fairly well(e.g., by showing that examinees with scores substantially above the passing score can solve most algebra items in the test and that examinees substantially below the passing score cannot do the algebra items on the test). Therefore, the descriptive assumption can be evaluated empirically by showing the passing score is or is not in more or less the right place on the score scale.

The policy assumption, which claims that the proposed performance standard is appropriate given the purposes of the decision process, adds a

second kind of arbitrariness to the standard-setting results, and is less amenable to empirical verification. This arbitrariness arises from the fact that decisions about how much is enough, like all policy decisions, are necessarily based on assumptions about the likely consequences of various choices and the values associated with these consequences, and although some of our assumptions about the consequences may be confirmed by data, assumntions about values are necessarily judgmental. If one group of judges decide that we should adopt competence in algebra as the standard on our math example because this will allow for a relatively rigorous treatment of content in the science course, and a second group wants to set the standard at a lower level, say competence in arithmetic, in order to allow more students to participate, it is hard to imagine any empirical study that could resolve this difference of opinion.

So, we have a fairly high degree of ambiguity or arbitrariness in our choice of passing score. In most cases, it appears that there is no specific passing score that can be considered the "correct" passing score, and as a result we cannot demonstrate that we have chosen the correct passing score. The best that we can do in "validating a performance standard" is to show that the passing score is consistent with the proposed performance standard and that this standard of performance represents a reasonable policy decision, given the overall goals of the assessment program. In practice, however, we seldom if every achieve even this goal. A more modest but realistic goal in most cases is to assemble evidence showing that the passing score is not unreasonable. In the next three sections, I will review three kinds of evidence used to achieve this goal.

To summarize the conclusions drawn in this section, we validate a performance standard by showing that the proposed interpretation for the passing score is reasonable and appropriate. The interpretation rests on two assumptions, a descriptive assumption claiming that the passing score corresponds to some performance standard, and a policy assumption claiming that the performance standard is appropriate, given the purposes of the decisions being made. It is possible, albeit difficult, to generate empirical evidence that supports the descriptive assumption directly. The policy assumption is not amenable to direct empirical verification, but some kinds of data can be relevant to the "reasonableness" of this assumption.

Procedural Evidence for Validity

Procedural evidence focuses on the appropriateness of the procedures used and the quality of the implementation of these procedures. As is true of most types of validity evidence, the procedural evidence can be more decisive in invalidating a standard than in validating it. Poor procedures or a failure to implement procedures in an appropriate way can destroy our confidence in the resulting passing score and performance standard. However, thorough implementation of the best available procedures does not guarantee that the resulting passing score is appropriate.

Procedural evidence is particularly important in evaluating the validity of performance standards for two reasons. First, in most cases, few if any

solid empirical checks on the validity of the performance standard are available. It is of course important to take full advantage of all opportunities for checking on the appropriateness of the standard and the associated passing score, and a number of methods for checking performance standards and passing scores are discussed in the two sections following this one, but given the severe limitations in the methods available, we are forced to rely heavily on procedural evidence. Second, procedural evidence is a widely accepted basis for evaluating policy decisions; we can have some confidence in standards if they have been set in a reasonable way (e.g. by vote or by consensus), by persons who are knowledgeable about the area for which the standards are being set, who understand the process they are using, who are considered unbiased, etc.

Selection of Procedures

As noted earlier, a number of procedures have been proposed for setting standards. Hambleton and Eignor (1980) reported on 18 standard-setting methods. Berk (1986) listed 38 separate procedures for either setting or statistically adjusting existing standards. As Berk (1986) notes, many of the procedures are variants on a few basic procedures, but the total number of options is still formidable. Unfortunately, we do not have any definite, authoritative guidelines for which procedures are to be preferred in general or in any particular case.

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985) impose a number of documentation requirements relevant to passing scores, but do not specify any particular procedures or types of procedures to be used. For example, Standard 6.9 requires documentation of the method used to set the standard and the rationale for this method. This standard also requires that the qualifications of judges be reported. Jaeger (1990, p. 15) interprets the requirement in Standard 3.1 that testing programs be developed on "a sound scientific basis" as implying that standard setting should "be well documented, be based on an explicable rationale, be public, be replicable, and be capable of producing a reliable result." Jaeger (1990, p. 16) notes that the Uniform Guidelines on Employee Selection Procedures, which apply to tests used for employment, add the requirement that the standard be "reasonable" and that the utility and adverse impact of the decision be examined and reported. Taken as a whole, these requirements do not eliminate any of the methods for standard setting describer earlier in this report, or for that matter any of the 38 methods listed by Berk (1986). All of these methods can be defended as being reasonable, and all can be made reliable if enough data are collected. Any procedure can be documented.

It is not surprising that the standards do not require or proscribe any method; the Standards were not intended to mandate or rule out specific methods in any area, but rather to provide general guidelines for good testing practices.

Piburn's (1990) analysis of legal challenges to licensure examination suggests the need for a "rational relationship" between requirements for

licensure and practice requirements. Piburn (1990, p. 14) quotes the decision of the Supreme Court in Schware v. Bd. of Bar Examiners, 1957, as follows:

> A State cannot exclude a person from the practice of law or from any other occupation in a manner or for reasons that contravene the Due Process or Equal Protection Clause of the Fourteenth Amendment.... A State can require high standards of qualification...but any qualification must have a rational connection with the applicant's fitness or capacity to practice [a licensed occupation].

Again, it seems that the standard embodied in the passing score needs to be reasonable or "rational", given the purpose of the decision process. However, Herbsleb, Sales, and Overcast (1985, p. 1169) suggest that the constitutional criteria for rationality is so lenient that the technical issues of validity are "simply irrelevant to the legal issues".

There is a substantial literature comparing the properties of various methods (e.g., Andrews and Hecht, 1976; Brennan and Lockwood, 1980; Koffler, 1980; Skakun and Kling, 1980; Mills, 1983; Cross, Impara, Frary and Jaeger, 1984), and much of this literature has been summarized by Jaeger (1989). This literature has provided us with insights into some of the problems and advantages associated with different methods; for example, the study by Brennan and Lockwood (1980), highlighted some potential problems with the Nedelsky procedure. Note, however, that Meskauskas and Norcini (1980) have suggested some advantages specific to the Nedelsky procedure. Collectively, these studies have also alerted us to the large differences in the passing scores that can occur when different methods are used, even with the same judges and same items, and have provided us with information about the reliability of different procedures.

However, this research has not decisively favored any one method over the others, for at least three reasons. First, and most fundamentally, most of these studies have compared the passing scores and passing rates for two or more methods, and found that one method produced a higher passing score than the other method. However, we have had no external criteria to indicate what the passing score should be. So we have had no way to decide which of the passing scores resulting from the different methods was to be preferred. The fact that the Angoff method yields a higher passing score than the Nedelsky method in a particular context, for example, does not tell us which of these two passing scores should be considered more appropriate. These studies have been extremely useful in several ways (e.g., in providing indications of the standard errors associated with various procedures), but because of the "criterion problem," they have not provided us with definite guidance on which method to use.

The two other reasons why this literature is hard to interpret in any simple way involve the variety of procedures that have been proposed and some inconsistencies in the results. Although there have been a number of studies comparing the different standard setting methods, these studies have involved different combinations of methods and variations on methods, different kinds of tests, and different contexts. So, the different studies are not

replications of eacn other, and it is difficult to draw general conclusions across studies.

The difficulty in identifying clear patterns in the results is increased by the fact that the results are not entirely consistent. For example, Brennan and Lockwood (1980) found that the standard error in estimating the passing score, over judges and items, was much smaller for the Angoff procedure than it was for the Nedelsky procedure, and they provided a possible explanation for this result in terms of the limited choices that the judges have in using the Nedelsky procedure. Smith and Smith (1988) also found that Angoff judges were more consistent than Nedelsky judges. However, a study by Cross, Impara, Frary, and Jaeger (1984) found that the Nedelsky method had a larger standard error than the Angoff method on a mathematics test and a smaller standard error for an elementary education test. Given the complexity of the comparisons to be made and the lack of complete consistency in the results, the conclusions to be drawn from this literature are not entirely clear.

The evidence seems to favor the Angoff procedure, but is not decisive, and therefore all of the proposed methods can be considered legitimate options. In practice, the Angoff method, in its many permutations, is most popular, and this is certainly not inconsistent with the literature. The Angoff method seems to be fairly convenient to use. It is flexible, allowing for gradual improvements in the specific procedures used, as possible limitations or problems with the method are identified (e.g., most users of the Angoff method now seem to provide judges with some data in order to provide feedback on consistency and possibly on the implications for pass rates of the decisions being made). Although the evidence is somewhat mixed, the Angoff method seems to have relatively small standard errors in the passing scores, and since reliability is a necessary condition for validity, the relative magnitudes of the standard errors is relevant to questions of validity. Note, however, that the reliability issue is not decisive because we can always decrease the standard error and therefore improve reliability for any method by increasing the number of judges, and/or items, and/or occasions used in standard setting.

Several studies have examined the relationship between the MPLs and item difficulty. Kane and Wilson (1984) showed that a positive covariance between item effects in judged MPLs, and item effects in examinee scores would lead to a reduction in the standard error of the passing score over samples of items and judges. Kane and Wilson (1984) also suggested that a positive covariance would support the validity of the judged MPLs, and a negative or zero covariance would suggest a lack of validity in these judgments, because a negative covariance would indicate that the item characteristics determining the MPLs were different from the characteristics influencing examinee performance. Halpin and Halpin (1987) found correlations between item difficulty and Ebel, Nedelsky, and Angoff MPLs for the Missouri College English Test of 0.49, 0.24, and 0.57, respectively. Busch and Jaeger (1990) found correlations between Angoff MPLs and item difficulties ranging from 0.30 to 0.78 over seven content areas; these correlations increased dramatically (ranging from 0.61 to 0.93) in a second round after the judges were given information on observed item difficulties, but the interpretation of these

higher correlations as checks on validity is complicated by the fact that the judges had been given data on item difficulty. Smith and Smith (1988) found a higher correlation between Angoff MPLs and p values (r = 0.60) than between Nedelsky MPLs and p values (r = 0.37), and suggested that this difference occurs because the Angoff ratings make use of more information that is relevant to item difficulty.

Berk (1986, p. 147) concludes that "...the Angoff Method appears to offer the best balance between technical adequacy and practicability." Cross, Impara, Frary, and Jaeger (1984, p. 126) concluded that, in their study comparing the Angoff, Nedelsky and Jaeger methods, the psychometric properties of the Angoff judgments "were unsurpassed". Smith and Smith (1988, p. 272) suggest that compared to the Nedelsky procedure, the Angoff procedure "encourages the use of a wider selection of information and information that is more predictive of item difficulty". Shepard recommends the Angoff and Jaeger methods as the, "most practical". However, Jaeger (1989, p 491) concludes that, "There is no agreement on a best method, although some procedures are far more popular than others." Gross (1985) has described the advantages of a modified version of the Nedelsky method. And of course, we have the view that we should not, in most cases set standards at all (Glass, 1978) because all of the methods are "blatantly arbitrary."

Psychometricians are not in complete agreement on the method of choice, but we do have some agreement on some issues. First, there are some situations (e.g., licensure and certification) where standard setting clearly seems necessary. There is consensus that in those cases where passing scores are to be used, they should be established in a careful and systematic way. There is some tendency among measurement specialists to prefer the Angoff and Jaeger methods over the Nedelsky method (e.g., Berk 1986; Brennan and Lockwood, 1980; Shepard, 1980), and the research on standard errors in passing scores and the relationships between items MPLs and item difficulties are consistent with this preference. The Angoff procedure seems to be the most popular method by far among those actually setting standards, suggesting that this method is found to be fairly convenient to use.

Implementation of Procedures

We have much more specific guidelines for the implementation of standard setting methods than we have for picking a method. Livingston and Zieky (1982) list five basic steps for any standard setting method: select the judges, define "borderline" knowledge and skill (equivalent to the performance standard, as defined here), train the judges in the use of the method chosen, collect judgements, and combine the judgements to choose a passing score. These criteria for evaluating implementation apply to all methods, although the details vary from one method to another.

In the remainder of this section, I will briefly discuss five parts of the standard setting procedures that have an impact on the plausibility of the standard, (1) definition of goals for decision procedure, (2) selection of judges, (3) training of judges, (4) definition of performance standard, and (5) data collection procedures. I will discuss these as they apply to the

content-centered standard setting methods. With some modification, these five criteria could be reworded to apply specifically to other types of standard setting studies.

(1) Definition of goals for decision procedure. The general purpose to be served by the use of a passing score needs to be defined before the standard-setting process, per se, begins. This general purpose is tied to the goals of the decision process as a whole, and as noted earlier, is often stated in terms of readiness for something (e.g., for licensure examinations, the purpose is to ensure "readiness" for entry-level practice). This general statement of purpose can be interpreted as a very general version of the intended performance standard. and provides a basis for the rest of the process.

(2) Selection of Judges. All of the standard setting methods involve judgements and therefore all need qualified judges. The specific qualifications needed in the judges will depend mainly on the type of decision to be made using the passing score (Jaeger, 1991). The technical expertise of the judges may be particularly critical in getting a good verbal description of the performance standard being developed. The familiarity of the judges with the population of examinees for whom the passing score will be used should help to keep the standard realistic.

Given the wide-ranging impact of the policy decisions involved in setting standards for high-stakes tests, it is probably important to have broad representation from groups with an interest in the stringency of the standard. However, this interest in having broad participation in the standard-setting process may be in conflict with the requirement that the judges be qualified to make the kind of decision they are being asked to make, and therefore, a judicious tradeoff may be called for. It is probably best not to restrict input to the standard-setting process to one group of experts (e.g., faculty in professional schools for licensure examinations, school teachers for high school exit examinations). Even though they may have less expertise in the area for which the standard is being set, it would generally be wise to include representatives from as many stakeholder groups as possible. However, each group should be asked to provide judgments in areas where they are qualified; it would be pointless to have a person with no knowledge of content make specific judgments about items.

In addition, the number of judges should be large enough, so that the standard error of measurement of the resulting passing score is not too large. The standard error can be evaluated in two ways, in terms of its magnitude and in terms of its impact on pass rates, and the standard error needs to be fairly small in both ways. We can generally decrease the size of the error by increasing the amount of data collected during standard setting. It is important to document the process used to estimate the standard error and the magnitude of the standard error. Note, standard errors of passing scores are discussed in more detail in the next section.

(3) Training of Judges. If they are to perform the task well, the judges need to understand what they are supposed to do. As a minimal amount of training, the judges should probably get an orientation to the goals of the

decision process and a detailed presentation on what they are to do during the rating process (Norcini, Lipner, Langdon, and Strecker, 1987; Mills, Melican, and Ahluwalia, 1991). Given that the judges may not have any experience in using standard-setting procedures, it would probably be reasonable to provide them with some practice and some feedback on their efforts (Reid, 1991), and periodic retraining if necessary (Plake, Melican, and Mills, 1991). Training should continue until both the judges and those conducting the study are satisfied that the judges understand what is expected of them.

(4) Definition of the Performance Standard. As part of the process of developing the standard, the judges are usually asked to develop a general definition of the standard of performance that they consider adequate for the intended purposes or goals of the decision process. Given this definition of the performance standard, the judges decide on the minimal passing levels for the items (in the test-centered models) or the Mastery/nonmastery status of the examinees (in the examinee-centered models). The performance standard is typically arrived at by consensus, even if the judges work independently in setting the passing score that is used to operationalize this performance standard.

The plausibility of the passing score resulting from this process is likely to be improved by having the judges take the time to reach agreement on a clearly stated definition of the performance standard being adopted. The rationale for the passing score is further strengthened if the judges can explicitly link the performance standard to the purpose of the decision to be made. For example, if the performance standard on a licensing or certification examination can be linked to the requirements of practice, the legitimacy of the standard is supported; if it can be shown that the detection of possible drug interactions is an important component of the safe practice of pharmacy, a standard that required pharmacists to know common drug interactions would be more defensible, and a passing score that implied that candidates for licensure had to be able to recognize potential interactions among commonly used drugs with a high degree of consistency would be more defensible than if this connection had not been explicitly made.

(5) Data Collection Procedures. Of course, the procedures used to collect the data need to be systematic and accurate. In addition to this basic requirement, there are some steps that are likely to improve the quality of the data.

First, if we want any work to be relatively error-free, it is generally necessary to review and/or check it at least once. Therefore, on this basis alone, iterative procedures in which the judges get to review their decisions before the passing score is finalized, seem to be preferable to single-pass procedures (Linn, 1978; Shepard, 1980; Jaeger, 1982, 1989; Busch and Jaeger, 1990). The multiple reviews incorporated in the iterative procedures are likely to be especially effective if they involve the introduction of new kinds of data, which are relevant to the standard setting task, at each iteration.

Second, it is highly desirable that the data collection procedures promote consistency in the data being generated. One way of doing this is to

have the judges discuss their ratings after they have independently judged the items. Fitzpatrick (1989) has pointed out some potential problems associated with group dynamics that we need to be concerned about, but the benefits of having the judges consider their judgements as a group seems to outweigh the risks. Statistical data on the performance of relevant groups of examinees can help the judges to set the passing scores at realistic levels (Jaeger, 1989; Hambleton and Powell, 1983; Shepard, 1980; Linn, 1978). In fact, Shepard (1980, p.463) argued that, "at a minimum, standard setting procedures should include a balancing of absolute judgments and direct attention to passing rates". Information on how the data for each judge compare to the data of other judges can also provide useful feedback.

If they are available, external checks on the reasonableness of the judgements being made would also be helpful. There is no good reason, in making a decision, to ignore information about the consequences of the decision, if such information is available (Linn, 1978; Busch and Jaeger, 1990; Norcini, Shea and Kanya, 1988). In many situations, we are forced to make decisions without knowing much about the consequences, but if good data on the probable consequences of decisions are available, it would seem to be prudent to use it and irresponsible to ignore it. Data on the consequences of setting the passing score at different points may be useful in helping judges to make realistic decisions. For example, judges, who are inclined to set a standard on a licensing exam that would either fail almost all recent graduates or pass all recent graduates, might be encouraged to reconsider their judgements.

Feedback from Judges

As an additional check on the design and implementation of the standard setting process, information can be collected from the judges about their perception of the process used to generate ratings, using rating forms or interviews. In particular, the judges are in a good position to provide information about their understanding of the purpose of the standard-setting study and the procedures that were used (Geisinger, 1991).

If the judges work from a predefined performance standard, the judges could be asked about their understanding of this performance standard. If the judges develop the performance standard, they could be asked about their understanding of the criteria used in developing the performance standard.

Finally, the judges could be asked about their level of satisfaction with the process as a whole and with the product, the resulting passing score. Did they think the process was sound and worked smoothly? Did they think that the conceptual definition of the standard was clear and appropriate? Did they think that the passing score was at an appropriate level and reflected the conceptual definition incorporated in the performance standard.

As with all of the different kinds of evidence for the validity of the passing score, a set of positive results obtained in a survey of the judges does not prove that the passing score is appropriate. All it says is that the individuals who developed the standard, think that it is appropriate. This is

not very strong evidence for validity. However, this kind of study can provide powerful evidence on the negative side. If the judges who developed the standard do not have confidence in it, it is not clear why anyone else should.

The fact that a standard setting study has employed an apparently sound procedure in a thorough and systematic way, and has where possible, included various checks on consistency and reality encourages us to have faith in the results. However, such procedural evidence does not provide strong assurance that the results are appropriate.

Evaluating Procedural Evidence

Procedural evidence cannot establish the appropriateness of a passing score and its associated performance standard, just as procedural evidence cannot establish the validity of a test score interpretation (Cronbach, 1971; Messick, 1989; Kane, 1992). However, procedural evidence can invalidate a standard, just as procedural evidence can invalidate a test-score interpretation. And procedural evidence can support the validity of a proposed interpretation of the passing score by ruling out one possible counterinterpretation. So procedural evidence is relevant to the validation of test score interpretations and standards.

Opinions vary on how important procedural evidence is in the validation of performance standards. Jaeger (1990, p. 18) suggests that:

> ...we examine the validity of a judgment-based standard-setting procedure by conceptualizing the universe score (Cronbach, Gleser, Nanda & Rajaratnam, 1972) that would result if ideal conditions of judgment were enjoyed by an ideal population of appropriate judges.

Jaeger seems to be endorsing the view that a good standard-setting study, one that approximates the "ideal", is the closest thing to a gold standard that we have. On the other hand, Madaus (1986, p. 13) has expressed some reservations about the utility of procedural evidence in supporting a proposed standard:

> Whether we like it or not we must face the reality that having 10 to 15 teachers make judgments about the percentage of examinees they think will pass an item, and using those judgments to arrive at a cut score, says nothing about the validity of any decisions made on the basis of the cut score. We need to go beyond a line of validity evidence resting on a partially rigged plebiscite to some empirical examination of the degree to which a test, using a particular cut score, in fact correctly separates those with insufficient knowledge and/or skills to teach at a minimally acceptable level from those who in fact have such prerequisite knowledge and/or skills.

Opinion is definitely mixed.

Perhaps the range of opinions on the efficacy of procedural evidence is so wide, in part, because procedural evidence plays very different roles in relation to our two assumptions. For the descriptive assumption, which claims that the passing score can be interpreted in terms of a specific performance standard, procedural evidence plays a role analogous to the role of procedural evidence in validating any score scale. Serious defects or omissions in the procedures can invalidate a proposed interpretation. Impeccable procedure provides support for the validity of the proposed interpretation by ruling out one possible counterinterpretation, but this support is quite limited.

Procedural evidence can play a much larger role in supporting the plausibility of the second assumption, which involves a policy decision. In our society, the legitimacy and defensibility of policy decisions are based to a large extent on procedural correctness. In most situations where important policy decisions are to be made, there are rules about who gets to vote, the place and time of the meeting, the number of voting members constituting a quorum, etc. If the rules are not followed, the policy decision is not considered legitimate, and if the rules are followed, the policy decision is accepted as legitimate. Individuals may not like the decision, they may think the decision to be foolish or unjust, but if the decision has been made in accordance with the rules, it has some legitimacy, just because the rules have been followed.

For example, if a duly constituted licensure board with the authority and responsibility for setting standards for entry to a profession, decides to set a particular passing score on the licensing examination, this judgment is likely to be accepted as a legitimate exercise of authority, unless compelling evidence indicating that this passing score is inappropriate is available. Werner (1978, p. 2) suggested that, in making judgments about standards, "an agency must weigh a variety of factors (both quantifiable and non-quantifiable) in making its final selection of a particular value." The responsible authority also has the right to change the standard; Ellwein, Glass, and Smith (1988, p. 22) commend decision-makers in South Carolina for passing 60% of the students who failed to achieve a performance standard: "We applaud the wisdom of those in South Carolina who interposed reason between crude technologies imposed on them from above and the lives of children."

The approach taken by the Florida Department of Education in setting passing scores on several high-stakes tests reflects this view of standard setting as policy making (Fisher, 1993). A committee chooses a passing score based on a review of the items using the Angoff procedure and a review of data on the performance of various groups. The committee recommendation goes to the Commissioner of Education and then to the State Board of Education. After these several reviews, the result is an administrative rule, with the weight of law. There is no attempt to collect external evidence of validity and no analyses of reliability are conducted (Fisher, 1993). Standard setting is treated as policy making, rather than as a technical problem of estimation.

Some policy-making body decides how strict or lenient the performance standard is to be, and whatever decision they make, they are likely to be commended by some and criticized by others (see Airasian, 1987). Mehrens (1986) and Madaus (1986) have provided an extremely interesting discussion of

the relative merits of certain policy options involved in standard-setting. The criteria applied in such debates tend to be procedural criteria (i.e., was the deed done properly) and general criteria of reasonableness.

The criteria for evaluating the legitimacy of policy decisions are clearly different from the criteria for evaluating the plausibility of scientific inferences. Claims that a score scale can be interpreted in terms of a certain kind of performance, or that a particular point on the scale can be interpreted in terms of a particular level of performance are amenable to direct empirical study, and are not accepted without empirical evidence (Cronbach, 1971, Messick, 1989). Policy decisions cannot be checked directly against data and their legitimacy is therefore evaluated in terms of general criteria of the reasonableness of the decision and the fairness, legitimacy, etc. of the procedures used to arrive at the decision. Therefore, procedural evidence can have a large impact on the plausibility of the policy assumption involved in standard-setting efforts.

A small and not particularly "scientific" survey of organizations engaged in setting standards for high-stakes achievement tests (e.g., licensure and certification tests, state mandated testing programs) was conducted in preparing this paper. I contacted ten organizations with responsibility for high-stakes achievement tests, including central testing agencies in four large states and six major professional organizations with responsibility for licensure and/or certification. I chose states and professional organizations that I thought likely to have engaged in studies of the validity of performance standards. I also contacted two major testing organizations and thereby obtained information on a large number of testing programs.

On the basis of this survey and a survey of the literature, the following generalizations seem relatively safe. First, the Angoff procedure in its various manifestations, is the standard method for setting passing scores on high-stakes achievement tests. Second, the details of the process vary quite a bit, but most organizations use an iterative procedure with some information on examinee performance provided at some point in the process. Third, most of these programs rely heavily on procedural evidence to support the validity of their standard setting procedures. In most cases, the only empirical checks on the validity of the standard setting focused on the reliability (i.e., standard errors) of the resulting passing scores. The most extensive work on the validity of the standards was that being done on licensure and certification tests in medicine (Norcini, 1993; Norcini and Shea, 1992; Norcini, Shea, and Kanya, 1988; Nungester, Dillon, Swanson, Orr, & Powell, 1991; Orr and Nungester, 1991).

Validity Checks Based on Internal Criteria

The data generated within the standard-setting study itself can be used as a partial check on the validity of the results. The emphasis in the design of internal checks is on the consistency of different sets of results derived from the study. Study results that are <u>not</u> internally consistent do not provide a solid basis for drawing any conclusions. Consistency in the results

does not provide compelling evidence for the validity of the proposed interpretation of the passing score, but it does provide some support for the passing score.

The internal validity checks are particularly relevant to the descriptive assumption. By checking various predictions based on the presumed relationship between the performance standard and the passing score, we check the claim that the passing score actually reflects the performance standard. The internal checks also provide indirect evidence for both assumptions by providing an empirical check on the consistency of the procedures used in the standard-setting study.

The Precision of Estimates of the Passing Score

No matter how well designed the standard-setting study and no matter how carefully implemented, we are not likely to have much faith in the outcome, if we know that results would be likely to be very different if the study were repeated. The extent to which we would be likely to get the same passing score if the study were repeated is indicated by the standard error of the passing score.

In order to estimate a standard error for the passing score in a meaningful way, we have to make some assumptions about the range of implementations of the standard-setting procedure that would be considered acceptable or exchangeable. Using the terminology of generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972), we need to define the intended universe of generalization. Presumably, different samples of judges could be used, and the data could be collected on different occasions, and therefore we would be willing to generalize over a judge facet and an occasion facet. We would also be likely to allow for some variations in how the standard-setting method is implemented in different studies (e.g., time spent on training may vary from study to study) and consider a "study" facet.

The item facet introduces some special complications. Most achievement tests assume that items are sampled from some domain of items, and passing scores are often generalized across different forms of a test using equating methodology. Nevertheless, the passing score is set for a particular set of items in the sense that for a given performance standard, the passing score should be lower for a difficult set of items than for an easy set of items. Therefore, in evaluating variability due to items, it is necessary to take account of the difficulty of the items (Kane and Wilson, 1984).

We can estimate the standard error in at least two ways. We can estimate the standard error directly by convening different groups of judges on two or more occasions, or two or more groups on the same occasion, and compare the results (Norcini and Shea, 1992). The disadvantage of this approach is that it is expensive to conduct multiple, independent standard-setting studies. The advantage of this design is that it provides us with a direct indication of how large the difference can be from one study to another using the same general design. It can also be applied to any kind of standard-setting study, including the Angoff, Nedelsky, Jaeger, and Ebel

procedures, contrasting groups, etc. If the standard-setting study employed the Angoff, Nedelsky, or Jaeger methods, we can also analyze the resulting data using generalizability theory (Cronbach, et al., 1972; Brennan, 1983) to obtain estimates of the variance components for judges, items, and studies. The variance component for studies would provide an indication of variability over occasions and variability due to differences in how the different studies were implemented. This is probably the best approach to estimating the standard error, but has not been used much because of its expense; it is necessary to conduct two or more independent standard-setting studies.

Alternately, if we have data from only one study involving the Angoff, Nedelsky, or Jaeger procedures, we can use generalizability theory to estimate the variance components for judges and items, and if data were collected on more than one occasion, we could estimate the variance component for occasions. The estimated variance components can then be combined to provide an estimate of the standard error in the passing score (Brennan and Lockwood, 1980; Kane and Wilson, 1984). This approach has the advantage of being easier to implement than the multiple-study approach. It has the disadvantage of not including all of the potential sources of error that are included in the dual-study approach. In particular, the estimate of the standard error doesn't include variability due to possible differences in the implementation across different studies.

The magnitude of the standard error should be evaluated against two criteria. To begin, the magnitude of the standard error can be evaluated by determining whether it is large compared to differences that are considered meaningful differences on the score scale. Meaningful differences on the score scale may be defined in terms of the magnitude of the standard error of measurement of examinee scores, particularly conditional standard errors near the passing score (Jaeger, 1991).

The second basis for evaluating the standard error of the passing score would be the variability in pass rates. In particular, we could examine how much the pass rate changes as we go from one standard error below the proposed passing score to one standard error above the passing score. A change of more than a few percentage points would probably be considered troublesome for high-stakes decisions.

Analysis of Item-level Data

There are at least two closely-related ways of using data from the standard-setting study along with data on patterns of examinee performance to provide useful checks on the passing score that is being chosen. Both of these methods could be incorporated in an iterative standard-setting procedure, with the data from earlier stage(s) being used in the subsequent stage(s) as checks on the results. Jaeger (1988) presents an alternative way of examining the consistency of standard-setting judgments.

The first method examines the relationship between item MPLs and the performance on the items for examinees with scores near the passing score (Kane, 1986, 1987). We can choose an interval around the passing score that

includes a fairly large number of examinee scores. The proportion of examinees in this interval answering an item correctly would provide an empirical estimate of the number of marginally competent examinees who can answer the item. This empirical estimate can then be compared to the MPL produced by the judges.

The results of such comparisons could be used as a check on the internal consistency of the ratings. To the extent that empirical estimates of the proportion of marginal examinees answering an item correctly differs from the original judgments about the probability of a minimally competent examinee answering the item correctly, there is some inconsistency in the results. Some differences are to be expected, but major inconsistencies would suggest a possible problem; for example, if items that the judges think almost all minimally competent examinees should be able to answer correctly are being answered by relatively low proportions of examinees with scores around the passing score, the passing score may be too low. Similarly, if examinees with scores around the passing score can generally answer items with low MPLs, the standard might be too high. In either case, we have some evidence that the item characteristics that the judges are using to evaluate item MPLs are different from the item characteristics that are determining the difficulty of the items for examinees, and this would cast doubt on the interpretability of resulting passing scores in terms of the performance standard.

The second method would involve the identification of two groups of examinees, one with scores a bit above the passing score and the other with scores a bit below the passing score. In this case we would expect the proportion correct on an item for the higher group to be above the MPL for the item and the proportion correct for the lower group to be below the MPL. Again, a confirmation of this expectation for most items tends to support the consistency of the results and a high proportion of failure to confirm tends to suggest inconsistency.

These analyses say something about the relationship between the performance standard and the passing score, especially if the analyses were consistent for a number of items. Assuming that the item MPLs were set on the basis of the proposed performance standard, and the performance standard is reflected in the passing score, the expected patterns in proportions correct should be found for most items. These analyses are essentially checks on the internal consistency of the process used to derive a passing score from the performance standard and therefore provide a check on the descriptive assumption. Confirmation of the expected findings does not say much about the appropriateness of the proposed standard, and therefore does not provide a direct check on the policy assumption.

Evaluating Internal Validity Checks

The internal checks on validity focus on the consistency of the results of the standard-setting study, in particular the consistency of the judges in translating the performance standard into a passing score. Therefore, they provide an empirical check mainly on the descriptive assumption, which posits a correspondence between the performance standard and the passing score.

The internal checks can also provide indirect support for the policy assumption by supporting the integrity of the procedures used to set the standard. However, because of their emphasis on the internal consistency of the judgments rather on the reasonableness of the judgments given the goals of the decision process, the internal checks do not provide a direct check on the appropriateness of the passing score.

Based on my small survey of organizations involved in high-stakes testing, and the content of the literature, it seems that most standard-setting efforts give some attention to the reliability of the results, either by examining interjudge reliability in some way or by computing a standard error for the passing score. However, the other possible internal validity checks are not generally used.

A good argument can be made for giving more attention to internal validity checks. The data needed for these analyses is relatively easy to collect, and can provide direct support for the descriptive assumption and for the integrity/consistency of the standard-setting process as a whole. To the extent that the results of the internal checks reveal problems (e.g., a judge or judges who are internally inconsistent and/or inconsistent with other judges), it may be possible to correct the problem before the passing score is finalized.

Validity Checks Based on External Criteria

A third type of validity evidence for the performance standard is based on comparisons with external sources of information about competence. In each case, we compare the results of decisions made using the passing score to the results of the same kind of decision or a related decision, made in a different way.

Of the two assumptions supporting the use of the passing score, these comparisons with external criteria tend to provide a check mainly on the policy assumption, which claims that the standard is appropriate given the purpose of the decisions. Each of these comparisons provides an indication, usually a "rough" indication, of whether the passing score is too high, too low, or about right. Some of these external comparisons also provide some information on the meaning of the performance standard and the meaning of the score scale, but these external validity checks are generally most relevant to the policy decision involved in adopting a particular level of stringency in the performance standard and its associated passing score.

There are many possible sources of data that could be used for this purpose, but none of these external checks on the validity of the proposed standard is definitive. A well-designed and carefully conducted standard-setting study is likely to provide as good an indication of the most appropriate passing score as any other source of information. There is no gold standard. There is not even a silver standard. The comparisons discussed in this section can be thought of as being analogous to convergent validity evidence for score scales (Campbell and Fiske, 1959). No single comparison is decisive, but a consistent pattern of results supporting the

appropriateness of the proposed passing score can provide convincing evidence for the standard, and a pattern suggesting that the proposed interpretation is inappropriate can provide convincing evidence against the standard.

The Direct, Criterion-related Approach

In many cases, passing scores are used to make decisions about "readiness" for some subsequent activity, such as further schooling, a job, professional practice. In such cases, the most direct way to examine the validity of the decisions would be to have a group of examinees complete the assessment and then have this group engage in the activity. If examinees with higher scores tend to do well in the activity and examinees with lower scores tend to do poorly, we have criterion-related validity evidence for the assessment results as a predictor of performance on the activity. If, in addition, pass rates are approximately equal for the assessment and the criterion performance of the activity, we have evidence that the passing score is appropriate.

This simple and direct approach has a lot of appeal, but it is hardly ever possible to implement in a completely satisfactory way for several reasons (Shimberg, 1981; Kane, 1985). First, it is usually not possible to develop a clearly valid measure of performance to use as a criterion measure; for example, how should one measure the quality of an individual's performance in professional practice. Second, in order to use this approach we need to define a performance standard on the criterion; this task is potentially more difficult than setting a passing score on the assessment. Third, in order to use this method, we need to evaluate how well examinees who pass and who fail the assessment perform at the activity. This is often impossible; it is unacceptable to allow individuals who have been judged unprepared to drive a car or treat patients to engage in these activities for a few months so that we can collect data for validity studies. As a result of these and other problems, the criterion-related approach is seldom used for high-stakes achievement tests.

Comparisons to Results of Other Standard-Setting Methods

One way to check on the appropriateness of the passing score resulting from a standard-setting study would be to conduct another standard-setting study on the same test using a different method (Werner, 1978). So if the Angoff method were used in the original study, the new study might involve the Nedelsky, Ebel, or Jaeger method. The comparison between the original passing score and the new passing score would provide an especially demanding empirical check on the appropriateness of the passing score, if it were implemented by different researchers, with a different group of judges, under different circumstances. Agreement between the passing scores derived from the two studies would provide support for the plausibility of the proposed passing score. Disagreement between the two studies casts doubt on the appropriateness of the proposed passing score, to the extent that the second study is considered to be as good or better than the initial study. However, in practical contexts, it is generally difficult to evaluate the quality of

the different standard-setting methods, and therefore the results of such comparisons are not decisive.

As noted earlier, there have been a number of studies comparing the results of different standard-setting methods in various contexts (Jaeger, 1989). In general, the different methods do not seem to be in close agreement in the passing scores that they generate. In studies comparing the Angoff, Nedelsky, Ebel, and Jaeger methods, the Nedelsky method tends to produce the lowest passing score, and the Ebel procedure tends to produce the highest passing score, but the number of studies that have examined each of the possible comparisons is not large and the results are not entirely consistent.

This kind of check is in a sense a foregone conclusion because the different methods have generally yielded substantially different results when they have been compared. In one sense, this finding is not very surprising because the different methods ask judges to attend to different aspects of items (e.g., the Nedelsky focuses on the distractors, the Ebel focuses on the content and difficulty of the item as a whole, and the Angoff focuses mainly on the difficulty of the items for marginal examinees) and to make different kinds of judgments (Scriven, 1978; Hambleton, 1978, Brennan and Lockwood, 1980; Meskauskas and Norcini, 1980). Nevertheless, if we consider the methods to be exchangeable in the sense that their results are interpreted in the same way, the disagreements are disturbing. As a solution to this problem, Hambleton (1978, p. 284) suggested that "considerable attention should be given to the selection of a method...and the implementation of that method". Scriven (1978, p. 274) suggested that the answer lay in better procedures. In the 15 years since Hambleton and Scriven wrote, some progress has been made in both directions. The procedures have been improved by making them iterative and by training judges more thoroughly and giving them additional information (e.g., on item difficulty levels) at various stages of the standard-setting process. There has also been a trend toward general use of some version of the Angoff procedure.

The usual way to compare two different standard-setting methods is to have the judges in the two studies start from the general goal of the decision procedure (e.g., to ensure that graduating twelfth graders have adequate mathematics skills), develop a more specific definition of the performance standard, and then set the passing score. Under these circumstances, a comparison of the results of the two studies provides an empirical check on the appropriateness of the proposed standard, given the goal of the decision procedure.

The obvious limitation in this kind of comparison is that the second study is not obviously better than the first. So, disagreement is hard to interpret. Agreement between the methods supports the appropriateness of the proposed standard, especially if the two studies were conducted independently with different sets of judges. Disagreement casts some doubt on the appropriateness of the standard, assuming that the appropriateness of the two methods being used is roughly equal.

An alternative design, in which a detailed description of the performance standard is provided to judges at the beginning of the study,

could focus on the clarity of the verbal description of the performance standard. The judges in the two studies would determine the passing scores for the same verbal description of the proposed performance standard (e.g., mastery of basic algebra). In this case, the comparison of the passing scores from the two studies would provide an empirical check on the clarity of the verbal description of the performance standard, but would not provide a check on the appropriateness of the standard.

Comparison to Pass/Fail Decisions Made with a Different Test

Scores for the same examinees from different tests can be used to check on the reasonableness of the proposed standard to the extent that the results of the second test are considered relevant to the decisions that need to be made. This approach can be especially attractive if the results of the second test already exist and are available, but it has the danger that the second test may be chosen more for convenience than for relevance.

If the second test covers the same kind of achievement as the test for which the standard is being set, e.g., mathematics, and if a comparable passing score has been set on the second test (comparable, in the sense that the passing scores were set to achieve the same general goal), the results of the two decision procedures can be directly compared to provide a fairly straightforward check on the comparability of the decisions made using the two procedures.

We can focus the analysis on the comparability of the passing scores by emphasizing consistency in pass rates. There are two ways in which examinees can get different results on the two tests; an examinee can fail the first test and pass the second test or an examinee can pass the first test and fail the second. I will refer to the first of these two kinds of combinations as a pass-fail pair and the second as a fail-pass pair. To the extent that the pass-fail rate is similar to the fail-pass rate, the two tests have standards that are comparable in their stringency. To the extent that they are different, the standards are different. By this criterion, the levels of the standards are considered different to the extent that the pass rates are different on the two tests.

This case provides a particularly graphic example of the general rule that none of the external checks on the validity of the standard is decisive, and, at best, each provides an indication of the appropriateness of the passing score. Note that two independently developed tests are not likely to cover the same content in the same way, and since the passing scores were presumably not set with exactly the same goal in mind, they cannot be expected to be exactly comparable. And, of course, there is usually no reason to think that the passing score on the second test is any better than the passing score under study. So, it is hard to interpret any differences in the results obtained using the two different tests as decisive evidence for or against the validity of the standard. Agreement between the results of the two procedures tends to support the appropriateness of the proposed standard, but does not prove it, and lack of agreement, especially a large discrepancy, suggests that

at least one of the two standards is inappropriate, but does not indicate which of the two standards is not appropriate.

In some cases, it might be reasonable to compare decisions based on assessments of two different areas of achievement, e.g., mathematics and language skills. For example, if many of the students in a school district who pass a basic skills test in language fail a basic skills test in mathematics, but very few of the students who pass the mathematics test fail the language test, and we have no reason to expect poorer performance in mathematics than in language skills, we might suspect that the passing score for the mathematics test is too high or the passing score for the language test is too low. Obviously, the difficulties in interpreting the results of such comparisons in any precise way are even more severe than they are if the tests cover the same area of achievement.

At best, a comparison of pass/fail decisions on tests covering two different areas of achievement provides a very rough check on the appropriateness of the proposed standard. Note that this kind of comparison assumes that we can expect similar levels of achievement in the two areas and that the standards in the two areas should be similar. These are both questionable assumptions in most cases.

Comparisons Involving Other Assessment Methods

The validity of the standard proposed for a test can be checked by comparing the pass/fail decisions made about a sample of examinees to the pass/fail decisions made about the same examinees using some other kind of assessment of the examinees' levels of achievement. The achievement level of each of the examinees might be assessed by an experienced teacher in a one-on-one assessment and, thereby, rated as being acceptable or not acceptable. As in the previous case, the critical comparison in the evaluation of the comparability of the two passing scores is the difference between the pass-fail rate and the fail-pass , or equivalently, between the pass rates using the two methods. If one has faith in the accuracy of the individual assessment, this comparison could be seen as providing evidence for or against the appropriateness of the proposed standard.

The different methods that have been proposed for setting standards could also be used in this kind of study. For example, assume that the proposed standard was developed using the Angoff method. In order to check on the appropriateness of the standard, one-on-one assessments or teacher judgment could be used to identify a marginal group of students who are considered to just meet the standard (see Livingston and Zieky, 1982, 1983). If these borderline students get scores on the test that cluster around the passing score, the appropriateness of the passing score is supported; if not, doubt is cast on the appropriateness of the passing score. Note that the force of this kind of check on validity depends on our ability to unambiguously identify borderline examinees, and it is not always clear how to do this.

A similar kind of study could be conducted by having the teachers identify one group of students whose level of achievement is clearly adequate and another group of students whose level of achievement is clearly inadequate. The first group might include examinees who have received instruction and the second group would be uninstructed (Berk, 1976; Werner, 1978; Hambleton, 1980). If most of the scores of the first group are above the passing score and most of the scores of the second group are below the passing score, the appropriateness of the passing score is supported. If most of the scores in both groups are above the passing score, or most of the scores in both groups are below the passing score, confidence in the appropriateness of the passing score decreases. This study would involve a comparison of the proposed standard to the results of the contrasting-group method (Livingston and Zieky, 1982).

I have assumed, in this discussion, that the passing score being checked was generated using the Angoff method, and that the borderline group method or the contrasting-group method is used to validate the proposed standard. Since the Angoff is the most commonly used standard-setting method, this is a reasonable way to talk about this kind of validity evidence. However, if the original standard were set using the borderline group method or the contrasting-group method, there would be no reason not to use the Angoff, Nedelsky, Ebel, or Jaeger method to provide an empirical check on the validity of the proposed standard.

Existing classification data could also be used as the basis for checking the appropriateness of the standard, if these existing classifications are relevant to the goals of the decision process for which a standard is being set. For example, suppose that the purpose of the test-based decisions is to identify students who already know enough of the content of a course that they can be given credit for the course without taking it and can possibly be placed in a higher level course (Frisbe, 1982; Willingham, 1974). In this context, it would probably be reasonable to assume that most students who have recently passed the course with a grade of "B" or better should pass the test and that most students who have never studied the topics covered in the course would not pass. The agreement between these expectations and the results of administering the test-based procedure would provide an indication of the appropriateness of the proposed passing score.

Comparisons of the decisions made using the proposed passing score to the pass/fail decisions made using some other assessment method, in particular comparisons of pass-fail to fail-pass rates, tend to indicate whether the standard is at roughly the same level for the two kinds of assessment. The appropriateness of the proposed standard is supported to the extent that the decisions made using it are consistent with other reasonable ways of making decisions about competence in an area.

Comparisons of Group Distributions

All of the external validity checks discussed up to this point have focused on analyses of individual scores. The appropriateness of the passing score can also be evaluated by analyzing information about distributions of

scores. In particular, the passing rates obtained using the proposed passing score can be compared to passing rates that have been found in other situations.

For example, if the distribution of achievement in the population of interest can be assumed to be at least roughly similar to the distribution in another population, and the proportion of individuals in that other population judged to be competent is known, we would expect a similar pass rate for the population of interest. If the pass rate obtained using the proposed passing score is close to the known pass rate in the other population, the appropriateness of the passing score is supported; otherwise, confidence in the appropriateness of the passing score decreases. So, for example, if the pass rate on a licensure examination has been 90% for a number of years, and during this period, new licensees have functioned satisfactorily in practice, and nothing has happened recently that would cause a sharp decrease in the competence of candidates for licensure, then the results of a new standard-setting study, conducted to obtain a passing score for a new form of the examination would be expected to yield a pass rate of approximately 90%. If the new passing score produced a pass rate of 60%, the appropriateness of the new passing score would be suspect.

In some cases, we might have expectations that certain populations should have very high pass rates or very low pass rates on an examination (Meskauskas and Norcini, 1980; Linn, 1978; Werrer, 1978). For example, we would probably expect the pass rate for a group of experienced, successful professionals in general practice to be very high if they were administered the licensing examination for their profession. On the other hand, we might expect a group of examinees with no education or experience related to the profession to have a low pass rate. If we were to administer the test to such groups and the pass rates using the proposed passing score were in agreement with these expectations, confidence in the appropriateness of the passing score would increase.

However, Jaeger (1990, p. 16) points out some of the difficulties inherent in getting any definite evaluation of the passing score out of this kind of data. As an example of the ambiguities in such comparisons, he suggests that in examining the "reasonableness" of a passing score used to screen applicants for initial teacher certification, we might assume that all qualified practicing teachers should pass. This certainly seems reasonable, but as Jaeger points out, the word "qualified" is not well defined. If we try to define qualified in terms of some percentile in the distribution of teacher scores, we end up making a fairly arbitrary selection. If we rely on peers or administrators to identify the qualified teachers, we have all of the problems associated with the use of relatively subjective and uncontrolled ratings. As with all of the external checks on validity, this check is probably most useful as a reality check. If a very high percentage of practicing teachers fail the examination using a particular passing score, the passing score is probably too high. So, we can potentially use this method to tell us when we are setting unreasonable standards, but we probably can't use such data to fine tune a standard.

Madaus (1986, p. 13) makes this kind of argument explicitly in relation to a competency test for beginning teachers, which only one candidate would have passed in the first administration. Madaus (1986, p. 13) argues that:

> Now on its face, given such a pass/fail rate, I personally don't think the original, unadjusted cut score was valid. (It's hard for me to believe that only a single individual from any of Alabama's fine institutions of higher education had the minimum knowledge and skills necessary to be a successful teacher.)

This kind of argument is convincing if and only if the results are extreme. As a result, such comparisons can be effective as reality checks, but are not very useful in selecting a particular passing score.

Judgments by Stakeholder Groups

Another source of data that can be used to evaluate the appropriateness of the standard is the judgments of groups with a strong interest in the outcomes of the decision process. In the case of school-based standards, these stakeholder groups might include parents, students, teachers, school board members, community leaders, etc. (Jaeger, 1982). For licensure and certification examinations, the stakeholder groups might include faculty in professional schools, public-interest groups, leaders in professional organizations, practicing professionals, the public, etc. (Orr and Nungester, 1991).

These stakeholder groups presumably do not study the assessment procedures in any detail in developing their evaluation of the appropriateness of the passing score, but rather use their experience, general knowledge, and judgment. This kind of evidence focuses almost exclusively on the policy question of how high the standard should be, i.e., on the appropriateness of the stringency of the passing score, and relatively little on the details of what is required of examinees, i.e., the definition of the performance standard. The collection of input from stakeholder groups is a time-honored part of democratic processes for establishing public policy, and can therefore be considered a reasonable way to support the appropriateness of the policy decision to set the standard at a particular level.

Evaluating the External Validity Checks

As noted earlier, it is highly unlikely that any single comparison would provide a definitive check on the appropriateness of the passing score. A large-scale, well-designed standard-setting study is likely to produce a standard and an associated passing score that are as plausible as the results of any other approach to setting a passing score. If a solid external criterion for the appropriate passing score (i.e., a "gold standard") were available, there would be no reason to conduct the kind of standard-setting study discussed in this paper. In any case where there is a disagreement

between two indicators of the appropriate passing score, either or both of the passing scores may be questioned.

If a series of external validity checks all suggest that the passing score is too high, it might be reasonable to conclude that the passing score is too high even if none of the comparisons being made is very decisive. Similarly, if a series of studies suggest that the passing score is too low, it might be reasonable to conclude that the passing score is too low. If the comparisons all show agreement with the proposed passing score, confidence in the reasonableness of this standard would increase. If the external checks were inconsistent, with some indicating that the proposed passing score is too high and some indicating that the proposed passing score is too low, it might be reasonable to accept the proposed passing score, but the inconsistency of the results would indicate that the standard of acceptable performance in the area is not clearly defined.

These external checks are generally most effective in evaluating the appropriateness of the general level of the standard and therefore are most relevant to the policy assumption. For example, in comparing performance on tests in different content areas, the most we are likely to get is a general indication of whether the standard is at an appropriate level (i.e., not too high and not too low). These external checks tend to provide relatively little evidence on the correspondence between the passing score and a particular definition of the performance standard.

External validity checks based on alternative standard-setting procedures have not been used much in practice, but there are some available examples. The pass/fail decisions for the certification examination of the National Board of Internal Medicine have been compared to ratings by the director of the program in which the candidates were trained. The program directors rate each candidate on a scale from 1 to 9. Candidates with ratings of 4 or 5, considered marginal, have a pass rate of around 50%, candidates with a rating of 9 had a pass rate of about 80-90% and candidates with ratings of 1, 2 or 3 had a pass rate of 20-30% (Norcini, 1993). The NBIM has also collected data from stakeholders (e.g., physicians, nurses) on whether certified practitioners perform better than uncertified practitioners (Norcini, 1993). The National Board of Medical Examiners has conducted studies that collected judgments of various stakeholder groups about the appropriateness of the passing score on the medical licensure examination (Orr and Nungester, 1991). Fabrey and Raymond (1987) surveyed recently certified nurses on whether the passing score on the certification examination was appropriate.

The relative rarity of external validity checks is probably due to the difficulty in collecting much of the data needed for such checks and the ambiguity of the results. As noted earlier, if the data needed for a decisive check on the validity of a passing score could be obtained with reasonable effort, it probably could be used to set the passing score in the first place. The Angoff and other standard-setting procedures are used because they are seen as being the most reasonable way to set a passing score. Therefore, the alternative sources of data that could be used in external checks on validity

are viewed as being at best comparable, and often inferior, to the procedures used to set the original passing score. As a result, a lack of correspondence between the original passing score and that suggested by the alternative approach is not compelling, and the evidence provided by the external validity checks tend to be highly ambiguous.

My small survey of current practice in standard setting indicates that external checks, based on judgments about the general reasonableness of the passing score, are routine. The results of standard-setting studies are generally not implemented in a mechanical fashion. The results are reviewed by the bodies with the ultimate authority for setting the passing scores, and it is not unusual for such bodies to modify the results (Mehrens, 1986) or, in some cases, to reject them altogether if they do not seem to be reasonable. Additional, intermediate reviews may be made by various committees with responsibility for the assessment process and/or by technical experts. Available data (e.g., passing rates from previous years, performance on other relevant measures in the same population or comparable populations) are likely to be considered. However, these reviews do not generally employ empirical checks on validity. Rather, they involve general reviews of the reasonableness of the results.

Given the potential weaknesses in all of standard-setting methods, and the responsibility of the decision makers to avoid making unreasonable decisions, multiple reviews of a proposed passing score seems to be appropriate. It is, of course, essential that any changes in the passing score not be made casually or capriciously, if the integrity of the process is to be preserved (Geisinger, 1991).

Conclusions

An analysis of the role of passing scores in interpretive arguments for high-stakes tests suggests that there are two assumptions that need to be evaluated in investigating the validity of a proposed interpretation of a passing score. Support for the descriptive assumption is to be derived mainly from procedural evidence and from internal validity checks. Support for the policy assumption is to be derived mainly from procedural evidence and external validity checks.

Procedural evidence is highly relevant to both assumptions, but serves different functions in the two cases. Procedural evidence can support the descriptive assumption by establishing a clear connection between the passing score and the performance standard. The procedural evidence can support the policy assumption by indicating that the policy decision, which focuses on how demanding the standard should be, has been made in an acceptable manner.

The descriptive assumption claims that the passing score can be interpreted in terms of a proposed performance standard. Examinees with scores above the passing score are assumed to have met the performance standard, in the sense that their levels of achievement are at or above the level of achievement specified in the performance standard. Examinees with scores below the passing score are assumed to have not met the standard.

There are several kinds of evidence that can support the descriptive assumption. This assumption tends to be more plausible to the extent that the standard-setting process involved an explicit statement of the performance standard and an explicit linking of the performance standard to the passing score. Internal checks on validity that relate scores above and below the passing score to different levels or patterns of performance on items or groups of items can strengthen our confidence in the correspondence between the passing score and the performance standard, by highlighting specific differences in performance between passing and failing examinees.

External checks on validity can also support the correspondence between the passing score and the performance standard, to the extent that the alternate decision process, to which the decisions based on the proposed passing score are compared in the external check, involves explicit criteria for evaluating performance. So, for example, if pass/fail decisions using the proposed passing score on a reading test are compared to direct assessments of reading skill that are based on explicit criteria of performance, the results may provide insight into the meaning of the passing score in terms of reading skill. However, if the alternate decision process involves global judgments of competence in reading, without any explicit performance criteria, the comparison will not relate the interpretation of the pass/fail decisions to a particular level of performance.

So, it is conceivable that procedural evidence, along with internal and external checks on the correspondence between the passing score and the performance standard could provide strong support for the descriptive assumption. If we have decided that examinees need to demonstrate a good working knowledge of algebra in some context (and even better if we have specified what we mean by a "good working knowledge of algebra" in some detail), we have a number of options for examining whether a passing score distinguishes students meeting this requirement from students not meeting the requirement. The correspondence between the passing score and the performance standard is largely an empirical question.

The policy assumption claims that the passing score and its associated performance standard are appropriate given the purposes of the decision process. This assumption is highly judgmental. It addresses the question of how much is enough, and in doing so incorporates values and assumptions about the consequences of various decisions. The choice of how demanding the standard should be is a matter of policy rather than an empirical question, and therefore cannot be answered empirically. Empirical studies can be helpful in informing and supporting such policy decisions, but do not in themselves imply a particular choice of how much is enough. Most of the inherent, unavoidable "arbitrariness" in standard setting resides in the policy decision associated with this assumption.

Procedural evidence can provide support for the appropriateness of a proposed standard, especially if the procedures used to set the standard are open to review and involve input from a wide sampling of individuals with an interest in the decisions being made and familiarity with the issues involved. That is, to provide support for the second assumption, the procedures to be

used are those that tend to lead to sound policy decisions and to generate public support for the policies.

Internal validity checks can also provide support for this assumption by suggesting that the procedures that were used were sensible and internally consistent. The internal checks provide only limited opportunities to directly evaluate the appropriateness of the standard because they focus on the internal consistency of the process used to set the standard. If these internal checks indicate that major inconsistencies occurred, we have reason to suspect the results. However, even perfect consistency in the process does not necessarily indicate the result was at an appropriate level; the judges could be in complete agreement in setting unreasonable standards. Therefore, it is probably the case that it is easier for these internal checks to undermine confidence in the appropriateness of the standard, than to provide strong support for appropriateness.

External checks on validity are especially relevant to claims about the appropriateness of the standard included in the policy assumption. Although any particular empirical check is likely to be inconclusive, a pattern of agreement between the results of the proposed decision process and other sources of information on competence supports the appropriateness of this passing score, and a pattern of disagreement with the results of other decision processes suggests that the passing score is inappropriate.

Although there are several kinds of evidence that are relevant to the second assumption, none of the methods for evaluating the appropriateness of the passing score make it possible to fine-tune the passing score. Rather, these methods provide reality checks which could be sensitive to major flaws in the performance standard, but would not be sensitive to small shifts in the standard. Therefore, even if we implement all available checks on the validity of the standard, the best that we are likely to be able to do is to show that the proposed standard is reasonable, or plausible.

As noted earlier, most of the evidence for the validity of passing scores is procedural evidence. Typically, some internal validity checks are implemented, in particular, evidence for the reliability of the results. External checks on the validity of passing score is rare, probably because of the difficulty of implementing the external checks and because of the likelihood that the results of the external checks will be highly ambiguous.

References

Airasian, P.W. (1987) State mandated testing and educational reform: Context and consequences. American Journal of Education, 95, 393-412.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). Educational Measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-references tests. Review of Educational Research, 56, 137-172.

Berk, R. (1976) Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 45, 4-9.

Block, J. H. (1978). Standards and criteria: A response. Journal of Educational Measurement, 15, 291-295.

Brennan, R. L. (1983). Elements of Generalizability Theory. Iowa City, IA: American College Testing.

Brennan, R. L. & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Burton, N. (1978) Societal standards. Journal of Educational Measurement, 15, 263-271.

Busch, J. C. and Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher examinations. Journal of Educational Measurement, 27, 145.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cronbach L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational Measurement, 2nd ed. (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley.

Cross, L., Impara, J., Frary, R. and Jaeger, R. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-130.

Ebel, R. L. (1972). Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall.

Ellwien, M.C., Glass, G.V., and Smith, M.L. (1988). Predilections, opinions, and prejudices. Educational Researcher, December, 21-22.

Fabrey, L. and Raymond, M. (1987). Congruence of standard setting methods for a nursing certification examination. Paper presented at annual meeting of National Council on Measurenment in Eduaction, Washington, D.C.

Fisher, T. (1993). Personal communication.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. Review of Educational Research, 59, 315-328.

Frisbie, D. A. (1982). Methods of evaluating course placement systems. Educational Evaluation and Policy Analyses, 4, 133-140.

Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. Educational Measurement, Issues and Practices, 10, 17-22.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Gross, L.J. (1985) Setting cutoff scores on credentialing examinations. Evaluation and the Health Professions, 8, 469-493.

Guion, R. M. (1974). Open a window: Validities and values in psychological measurement. American Psychologist, 29, 287-296.

Halpin, G. and Halpin, G. (1987) An analysis of the reliability and validity of procedures for setting minimum competency standards. Educational and Psychological Measurement, 47, 977-983.

Hambleton, R. K. (1978). On the use of cutoff scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

Hambleton, R. F. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.) Criterion-referenced measurement, the state of the art. Baltimore, Johns Hopkins University Press.

Hambleton, R. K., & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum Competency Achievement Testing: Motives, Models, Measures, and Consequences (pp. 367-396). Berkeley, CA: McCutchan.

Hambleton, R. and Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-references testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Herbsleb, J. D., Sales, B. D., and Overcast, T. D. (1985). Challenging Licensure and Certification. American Psychologist, 40, 1165-1178.

Jaeger, R. M. (1979). Measurement consequences of selected standard-setting models. In M. A. Bunda & J. R. Sanders (Eds.), Practices and Problems in Competency-based Measurement. Washington, DC: National Council on Measurement in Education, 1979.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-476.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgements. Applied Measurement in Education, 1, 17-31.

Jaeger, R. M. (1989). Certification of Student Competence (pp. 485-514). In R. L. Linn (Ed.), Educational Measurement, 3rd ed. New York: American Council on Education and Macmillan.

Jaeger, R. M. (1990). Establishing standards for teacher certification tests. Educational Measurement, Issues and Practices, 9, 15-20.

Jaeger, R. M. (1991). Selection of judges for standard setting. Educational Measurement, Issues and Practices, 10, 3-6, 10, 14.

Kane, M. (1987) On the use of IRT models with judgmental standard-setting procedures. Journal of Educational Measurement, 24, 333-345.

Kane, M. (1985) Definitions and Strategies for Validating Licensure Examinations. In J. Fortune (Ed.) Understanding Testing in Occupational Licensing (pp. 45-64) Jossey-Bass, San Francisco.

Kane, M (1986). The interpretability of passing scores. ACT Technical Bulletin, Number 52, ACT, Iowa City.

Kane, M. and Wilson, J. (1984) Errors of measurement and standard setting in mastery testing. Applied Psychological Measurement, 8, 107-115.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Levin, M. (1978) Educational performance standards: Image or substance? Journal of Educational Measurement, 15, 309-319.

Linn, R. L. (1979). Issues of validity in measurement for competency-based programs. In M. A. Bunda & J. R. Saunders (Eds.), Practices and Problems in Competency-based Measurement. Washington, DC: National Council on Measurement in Education.

Linn, R.L. (1978). Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 15, 301-308.

Livingston, S. A. & Zieky, M. J. (1982). Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, NJ: Educational Testing Service.

Livingston, S. A. & Zieky, M. J. (1983). A Comparative Study of Standard-setting Methods (Research Report No. 83-38). Princeton, NJ: Educational Testing Service.

Madaus, G. F. (1983). Minimum competency testing for certification: The evolution and evaluation of test validity. In G. F. Madaus (Ed.), The Courts, Validity, and Minimum Competency Testing. Hingham, MA: Kluwer-Nijhoff.

Madaus, G. F. (1986) Measurement specialists: testing the faith - a reply to Mehrens. Educational Measurement, Issues and Practice, 5, 11-14.

Mehrens, W. A. (1986) Measurement specialists: motive to achieve or motive to avoid failure? Educational Measurement, Issues and Practice, 5, 5-10.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 45, 133-158.

Mesdauskas, J. & Norcini, J. (1980) Standard-setting in written and interactive (oral) specialty certification examinations. Evaluation and the Health Professions, 3, 321-360.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. Educational Researcher, 10, 9-20.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer and H. Braun (Eds.), Test Validity (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement, 3rd ed. (pp. 13-103.) New York: American Council on Education and Macmillan.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

Mills, C. N., Melican, G. J., and Ahluwalia, N. T. (1991). Defining minimal competence. Educational Measurement, Issues and Practices, 10, 7-9.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Norcini, J. (1993). Personal communication.

Norcini, J., Shea, J., and Kanya, D. (1988). The effect of various factors on standard setting. Journal of Educational Measurement, 25, 57-65.

Norcini, J., Lipner, R., Langden, L. and Shecker, C. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Norcini, J. and Shea, J. (1992). The reproducibility of standards over groups and occasions. Applied Measurement in Education, 5, 63-72.

Nungester, R. J., Dillon, G. F., Swanson, D. B., Orr, N. A., & Powell, R. D. (1991). Standard-setting Plans for the NBME Comprehensive Part I and Part II Examinations. Academic Medicine, 66, 429-433.

Orr, N.A., & Nungester, R. J. (1991). Assessment of Constituency Opinion about NBME Examination Standards. Academic Medicine, 66, 465-470.

Piburn, K. M. (1990). Legal Challenges to Licensing Examinations. Educational Measurement, Issues and Practices. NCME, 9, 5-6.

Popham, W. J. (1978). As always provocative. Journal of Educational Measurement, 15, 297-300.

Plake, B. S., Melican, G. L., and Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. Educational Measurement, Issues and Practices, 10, 15-16, 22, 25.

Reid, J. B. (1991). Training judges to generate standard-setting data. Educational Measurement, Issues and Practices, 10, 11-14.

Scriven, M. (1978). How to anchor standards. Journal of Educational Measurement, 15, 253-275.

Shepard, L. A. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), Practices and Problems in Competency-based Measurement. Washington, DC: National Council on Measurement in Education.

Shepard, L. (1980) Standard setting, Issues and methods. Applied Psychological Measurement, 4, 447-467.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), A Guide to Criterion-referenced Test Construction (pp. 169-198). Baltimore: Johns Hopkins University Press.

Shimberg, B. (1981) Testing for licensure and certification. American Psychologist, 36, 1138-1146.

Skakun, E. N. & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

Smith, R.L. and Smith, J.K. (1988) Differential use of item information by judges using Angoff and Nedelsky procedures. Journal of Educational Measurement, 25, 259-274.

van der Linden, W. J., & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1, 593-599.

Werner, E. (1978) Cutting scores for occupational licensing tests, manual of considerations and methods, California Department of Consumer Affairs, Sacramento.

Willingham, W. (1974). College Placement and Exemption. New York: College Entrance and Examination Board.